Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation

Anonymous ACL submission

Abstract

Personalized dialogue systems explore the problem of generating responses that are consistent with the user's personality, which have raised much attention in recent years. Existing personalized dialogue systems have tried to extract user profiles from dialogue history to guide personalized response generation. Since the dialogue history is usually long and noisy, most existing methods truncate the dialogue history to model the user personality. Such methods can generate some personalized responses, but a large part of dialogue history is wasted, leading to sub-optimal performance of personalized response generation. In this work, we propose to refine the user dialogue history from a large scale, based on which we can handle more dialogue history and obtain a more abundant and accurate persona information. Specifically, we design an MSP model which consists of three personal information refiners and a personalized response generator. With these multi-level refiners, we can sparsely extract the most valuable information (tokens) from the dialogue history and leverage other similar users' data to enhance the personalization. Experimental results on two real-world datasets demonstrate the superiority of our model in generating more informative and personalized responses.[1]

1 Introduction

Recent years have witnessed great progress on building personalized dialogue systems. In general, previous work explores building a personalized dialogue system via two pathways: (1) directly modelling user personality from predefined persona descriptions or user attribute (Qian et al., 2018; Zhang et al., 2018; Song et al., 2019); and (2) implicitly modelling the user personality from the user's dialogue history (Li et al., 2016c; Ma et al., 2021). The latter is considered superior as the dialogue history is easy to obtain and comprises rich personalized information. In this paper, we follow the second pathway that automatically learns implicit user profile from the user's dialogue history to assist personalized response generation.

It is challenging to model user personality directly from the dialogue history. The reason is that a user's dialogue history might contain massive historical dialogues, which are too heavy to load in the model and likely to be noisy. A straightforward solution is to truncate the dialogue history, as has been done by existing work (Ma et al., 2021; Qian et al., 2021a). However, as tremendous information has been wasted, the model's performance is also influenced. On the other hand, we observe that the dialogue history from other users' may also be helpful in generating a personalized response for the current user. For example, users with the same interest on "soccer" may talk about similar things on such a topic. This has been overlooked by existing methods. Intuitively, the problem of "data explosion" is even more severe in the latter case (when other similar users' dialogue history is also considered). To alleviate these problems, we propose using a hierarchical refiner structure to sparsely extract the most valuable query-aware persona information from both the current and other similar users' dialogue history. By this means, more dialogue history can be utilized for learning user personality and improving response generation.

Our model is called **MSP**, which stands for **M**odeling and **S**electing user **P**ersonality from the dialogue history for generating personalized responses. Instead of attending to all dialogue history, MSP refines the most valuable historical information that can well portray the user personality and guide the response generation. Specifically, MSP consists of three personal information refiners working in different levels and a personalized response generator. At **first**, a user refiner is designed to select a group of users who have similar interests with the current user. By refining dialogue

[1] Our codes are released in anonymous.4open.science.

history at the user level, we can obtain similar data to share information with similar users and avoid mutual interference with other users. **Then**, a topic refiner filters out the current and similar users' dialogue history that has different topics with the current query in sentence level. **Next**, we design a token refiner to extract the most valuable query-aware user profiles from the remaining dialogue history in the token level. **Finally**, a personalized response generator combines user profiles and the current query to generate responses. Given that there is no explicit supervisory signal guiding the refiner extract an exemplary user profile, we design a supplementary sentence matching task and a joint training method. The generator will construct a pseudo-label to guide the refiner's extraction.

Our contributions are three-fold: (1) We design an MSP model to tackle the data noise problem. It can efficiently refine user profiles through dialogue history and generate personalized responses. (2) We design a refiner structure to extract the query-aware profile in three levels. The similar users' information is taken into account, which can help improve the personality for the response. (3) We design a joint training method of the refiner and generator. The refiner provides the generator with user profiles to assist in generating responses, while the generator constructs a pseudo-label for the refiner to assist in selecting user profiles.

2 Related Work

Personalized dialogue generation. Open-domain dialogue generation has been extensively studied (Koehn et al., 2003; Vinyals and Le, 2015; Serban et al., 2016; Zhang et al., 2019a,b; Xiao et al., 2020). Recently, personalized dialogue systems have attached more and more attention. Typical methods include: (1) explicitly using predefined persona descriptions or attributes as users' profile to generate personalized responses (Qian et al., 2018; Zhang et al., 2018; Olabiyi et al., 2019; Song et al., 2019); (2) using user ID embeddings to enhance personalized dialogue generation (Li et al., 2016c; Chan et al., 2019); and (3) extracting implicit user profile from users' dialogue history to generate personalized responses (Al-Rfou et al., 2016; Bak and Oh, 2019). Since manually collecting user profiles is impractical for large-scale datasets and the user ID embeddings perform badly, in this study, we focus on the last group of methods for personalized response generation.

DHAP (Ma et al., 2021) is the state-of-the-art method in personalized dialogue generation. It uses a transformer-based structure to model the user's dialogue history and extract personal information for response generation. Unfortunately, this model can only handle a limited number of user dialogue histories, wasting a lot of valuable information. Our method has two main differences with DHAP: (1) We propose a refiner structure in our model so that more dialogue history can be handled and the most valuable information can be extracted for improving response generation. (2) With our proposed refiner, we can further incorporate more dialogue history from other users (having similar interests) to facilitate personalized dialogue generation.

Retrieval-guided natural language generation. Retrieval-based methods can collect relevant information for language generation (Yang et al., 2019). It has been widely applied in many tasks such as text style transfer (Li et al., 2018) and dialogue generation (Wu et al., 2019; Cai et al., 2019). The idea of using a retrieval system to get useful information inspires our study. We use a refiner to automatically extract personal information from dialogue history and guide the personalized generation.

3 Methodology

3.1 Problem Statement and Overview

Considering a set of users $\mathcal{U} = \{u_1, \cdots, u_l\}$, for any user u_i, we have their dialogue history with others $U^i = \{(q_1^i, r_1^i), \cdots, (q_t^i, r_t^i)\}$, where q_j^i is a *query* issued by others, while r_j^i is the corresponding *response* given by u_i.[2] Our target is to generate a personalized response r^i for the user u_i to reply a new query q. As we introduced earlier, the personalized information can be obtained from the dialogue history U^i of the user u_i and other dialogue history U^j from similar users $u_j (j \neq i)$.

The overview of our MSP model is shown in Figure 1. MSP consists of three personal information refiners working in different levels and a personalized response generator. Specifically, the first refiner is working at the user-level. By comparing the dialogue history of the current user u_i with others, MSP can select a group of users having similar interests with u_i. After obtaining a group of similar users, we further refine their dialogue

[2]Here we use the term "query" to denote the utterance given by others. Generally, the query can be either one utterance in single-turn dialogues, or several history utterances in multi-turn dialogues.

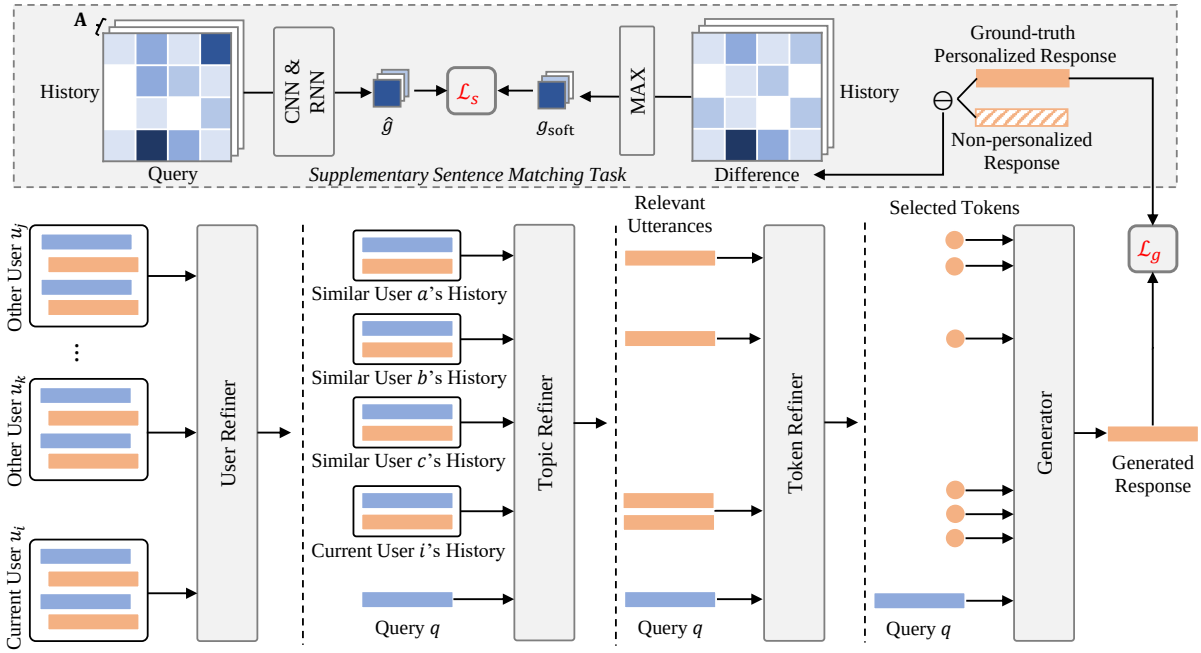

Figure 1: The overview structure of the proposed model which consists of four modules: (1) user refiner, (2) topic refiner, (3) token refiner, and (4) generator.

history according to the relevance with the current query's topic. Moreover, we add the last refiner to extract several tokens so that the most fine-grained personal information can be extracted from the relevant utterances. Finally, the query and the extracted tokens are fed into the generator and construct a personalized response.

3.2 User Refiner

The dialogue history of users with similar interests may share much personal information. Therefore, our first target is to select a group of users with similar interests to the current user. We design a user refiner to achieve this. Since the users' interest is usually contained in their dialogues with others, we consider both the queries and responses in the dialogue history to select similar users. Specifically, for the user u_i's dialogue history U^i, we apply a pre-trained BERT (Devlin et al., 2019) and represent them by the embedding of "[CLS]" token:

$$\mathbf{U}_q^i = \sum_{j=1}^{t} \mathrm{BERT}(q_j^i), \quad \mathbf{U}_r^i = \sum_{j=1}^{t} \mathrm{BERT}(r_j^i).$$

Then, we can select k_u users that have similar interest with the current user u_i:

$$u^{\mathrm{sim}} = \mathrm{TopK}(\mathbf{U}^i \cdot \mathbf{U}^j, k_u), \quad (1)$$

$$\mathbf{U}^i = [\mathbf{U}_q^i; \mathbf{U}_r^i], \quad (2)$$

where $\mathrm{TopK}(\cdot,\cdot)$ is the top-k selection operation.

After the user refiner, we can obtain the dialogue history of the similar users $\{u_j\}_{j=1}^{k_u}$. It is worth noting that, since the number of users is large in the datasets, we choose to use the dot-product to compute the similarity of the users so that the whole process can be implemented by dense retrieval libraries, such as Faiss (Johnson et al., 2017), which is very efficient.

3.3 Topic Refiner

The users' dialogue history often contains many dialogues with others. These dialogues have various topics, which may be irrelevant to the current query. Therefore, we propose a topic refiner to select relevant dialogue history for personalized response generation. Specifically, we use a topic classifier to compute the topic distribution of the current query q and the queries q_j^i in the history dialogues:

$$t = \mathrm{MLP}(\mathrm{mean}(\mathrm{BERT}(q)), \quad (3)$$

$$t_j^i = \mathrm{MLP}(\mathrm{mean}(\mathrm{BERT}(q_j^i)), \quad (4)$$

where $t, t_j^i \in \mathbb{R}^{d^t \times 1}$, and d^t is the number of topic. Then, we filter out the dialogue history $< q_j^i, r_j^i >$ that has different topics with the current query, *i.e.*, $\max(t_j^i) \neq \max(t)$.

In the topic refining process, we compare the queries in the history dialogues with the current query to filter out topic-irrelevant dialogues. This

process can further reduce the noise and make our model more lightweight. Both the dialogue history of the current user and that of the similar users (obtained in the former step) are refined. In the next step, we will use the **responses** in these selected history dialogues and extract the most valuable tokens for personalized response generation.

3.4 Token Refiner

After the previous two refiners, we obtain a collection of historical responses. Though we can directly add them into the generation process, our preliminary experiments indicate that they perform poorly. A major reason is the noisy quality of the responses. Indeed, existing studies (Shang et al., 2015; Borgeaud et al., 2021) have demonstrated the effectiveness of using informative tokens to improve the response generation. Inspired by these studies, we further devise a token refiner to extract the most fine-grained information (tokens) from the historical responses. Specifically, we compute an attention map $\mathbf{A}$ between the query q and the historical responses r^{sim} and r^{cur} (they are from the similar users and the current user respectively) as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right), \tag{5}$$

$$\mathbf{Q} = \text{TRM}_{\text{enc}}(q) \cdot \mathbf{W}_Q, \tag{6}$$

$$\mathbf{K} = \text{TRM}_{\text{enc}}(r) \cdot \mathbf{W}_K, \tag{7}$$

where $\text{TRM}_{\text{enc}}(\cdot)$ is a transformer encoder. r refers to r^{sim} or r^{cur}, and correspondingly, $\mathbf{A}$ refers to the similar user matching map $\mathbf{A}^{\text{sim}}$ and current user matching map $\mathbf{A}^{\text{cur}}$. $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d\times d}$ are parameters, and d is the dimension of the hidden state. After obtaining the attention matching map $\mathbf{A}$, we select tokens to form the similar users' profile and current user's profile according to each token's attention weight:

$$\mathbf{c}^{sim} = \text{TopK}(\text{Max}(\mathbf{A}^{\text{sim}}), k_p), \tag{8}$$

$$\mathbf{c}^{per} = \text{TopK}(\text{Max}(\mathbf{A}^{\text{cur}}), k_p), \tag{9}$$

where k_p is a hyper-parameter to control the number of profile tokens.

3.5 Generator

We use a transformer decoder as to generate a personalized response by using the similar users' profile $\mathbf{c}^{\text{sim}}$, current user's profile $\mathbf{c}^{\text{cur}}$, and query information $\mathbf{q}$ as input. The decoding process can be defined as:

$$\hat{y} = \text{TRM}_{\text{dec}}(\mathbf{x}), \tag{10}$$

$$\mathbf{x} = [\mathbf{c}^{\text{sim}}; \mathbf{c}^{\text{per}}; \mathbf{q}], \tag{11}$$

where [;] is the concatenation operation and $\hat{y}$ is the word generation probability.

3.6 Training and Optimization

The generator is optimized by maximizing the generation probability of the ground-truth y:

$$\mathcal{L}^{\text{g}} = -y \log \hat{y}. \tag{12}$$

In practice, we find that the token refiner is hard to train. We speculate the reason is a missing of direct supervision signals. In this case, it is difficult to tell whether the training errors stem from the generation process or the refining process. To tackle this problem, we propose a supplementary sentence matching task to assist the token selection.

Supplementary sentence matching task. The core idea of this task is to train the token refiner directly by introducing supervision signals so that it can automatically mine valuable tokens. Specifically, we design a sentence matching task to match the query with the dialogue history. The task's target is to find the history sentences that help generate personalized responses. We consider using the query-history cross attention weight $\mathbf{A}$ to generate a matching representation and then use this representation to finish the task. In this way, once the matching task has been well-finished, we can use attention map $\mathbf{A}$ to identify the most informative and valuable parts of a history sentence that are helpful to generate a personalized response.

To achieve our idea, we design a matching process. Firstly, we calculate the matching representations $\mathbf{H}$ by the cross-attention map $\mathbf{A}$ and then apply a CNN with a max-pooling operation to aggregate token information:

$$\mathbf{S} = \text{Maxpool}(\text{CNN}(\mathbf{H})) \tag{13}$$

$$\mathbf{H} = \mathbf{A} \cdot \mathbf{V}, \tag{14}$$

$$\mathbf{V} = \text{TRM}_{\text{enc}}(r) \cdot \mathbf{W}_V. \tag{15}$$

Next we flatten $\mathbf{S}$ and applies a LSTM to aggregate the sentence information:

$$\mathbf{h} = \text{LSTM}(\text{Flatten}(\mathbf{S})). \tag{16}$$

Finally we use sentence matching vector $\mathbf{h}$ to compute the matching score:

$$\hat{g} = \text{Sigmoid}(\text{MLP}(\mathbf{h})). \tag{17}$$

Algorithm 1 Joint training process

Input: M dialogue triplets:$D = \{\langle q_i, y_i, r_i \rangle\}_{i=1}^{M}$
Output: A personalized dialogue model

```
Init the refiner and generator module
while not converge do
    Sample n_s dialogue triplets D^q = {q_i, y_i, r_i}_{i=1}^{n_s}
    Get Ŷ^q = {ŷ_i}_{i=1}^{n_s} on D^q from p_g(ŷ|q)
    Get pseudo-label g on D^q
    Train refiner by optimizing L^s on D^q
    if Current Step > N_f then
        Sample n_d dialogue triplets D^p = {q_i, y_i, r_i}_{i=1}^{n_d}
        Extract C^p = {c_i}_{i=1}^{n_d} on D^p from p_s(c|q, r)
        Train generator by optimizing L^g on D^p ∪ C^p
    end if
end while
```

For guiding the sentence matching task, we design a pseudo-label g to measure the matching goodness of each history sentence. We expect the history with more persona profile information can achieve a higher score. Thus we use the difference between personalized ground-truth y and non-personalized generated response probability $\hat{y}$ to measure the persona profile and create the pseudo-label:

$$g = \begin{cases} 1, & g_{soft} \geq \alpha \\ 0, & g_{soft} < \alpha, \end{cases} \tag{18}$$

$$g_{soft} = \mathrm{Sum}(\mathrm{Max}((y - \hat{y}) \cdot r))/d_{\mathrm{y}}, \tag{19}$$

where d_{y} is the length of ground-truth y and α is a threshold. Finally, we minimize the binary cross entropy loss between g and $\hat{g}$:

$$\mathcal{L}^{\mathrm{s}} = g \log \hat{g} + (1 - g) \log(1 - \hat{g}). \tag{20}$$

Joint training. To facilitate the learning with the above gradient approximation approach, we design a joint training process to train the refiner and generator in turn. Specifically, in each training iteration, we first sample a batch of query q, response y, similar and current users' dialogue history r^{sim} and r^{per} from dataset D. After through a non-personalized generator, we create the pseudo-label g (Eq.18). This pseudo-label is used to train the token refiner by optimizing the loss $\mathcal{L}^{\mathrm{s}}$ (Eq. 20). Further, we sample another batch D^p from D. After extracting sim profile $\mathbf{c}^{\mathrm{sim}}$ and persona profile $\mathbf{c}^{\mathrm{per}}$ (Eq.8, Eq.9), we generate the personalized response $\hat{y}$ and update the generator by optimizing the loss $\mathcal{L}^{\mathrm{g}}$ (Eq. 12). To avoid bad profile misleading the generation at the beginning training process, we pre-train the refiner for N_f steps before extracting the profile for the generator. The detailed training process is summarized in Algorithm 1.

4 Experiments

4.1 Datasets

To evaluate our model's performance, we conduct experiments on a Chinese Weibo[3] dataset (Qian et al., 2021b) and an English Reddit[4] dataset. Both are collected from open-domain social media platforms. On these platforms, users can post various topics, and other users can respond to them. We compare user-id and timestamps to associate the query with its corresponding response and the current user's dialogue history. As a result, each training sample contains a query, a response, and a sequence of dialogue history. Finally, the dataset is divided into training, validation, and test sets in chronological order. The statistics of the datasets are provided in Appendix A.

4.2 Baselines

We compare our proposed model with eight highly correlated and strong baselines. [5] They can be categorized into four groups:

Non-personalized methods. (1) Seq2Seq-Attention (Sutskever et al., 2014) is a vanilla sequence-to-sequence model with attention mechanism (Luong et al., 2015). (2) MMI (Li et al., 2016a) is based on seq2seq and use maximum mutual information as an extra loss to improve diversity. (3) DialoGPT (Zhang et al., 2019b) is a variant of GPT-2 (Radford et al., 2019) designed for dialogue generation.

Predefined profile-based methods. Since there are no persona descriptions in the datasets, we test these methods by using the user's dialogue history as a simulation of predefined persona profiles. (4) GPMN (Zhang et al., 2018) enhances the seq2seq model with a memory module, which encodes and stores the persona profile as memory representations. (5) PerCVAE (Zhao et al., 2017) encodes predefined personalized sentences as a conditional representation and uses CVAE to generate a personalized response.

User ID-based methods. (6) Speaker (Li et al., 2016c) is based on seq2seq while using user-id embedding as user representation to facilitate the response generation. (7) Persona WAE (Chan et al., 2019) uses WAE (Wasserstein autoencoder) for response generation. It maps user-id embeddings to a personalized Gaussian mixture distribution and

[3] https://www.weibo.com/
[4] https://www.reddit.com/
[5] The implementation details are given in Appendix B.

then samples the personalized vector to guide the response generation.

User dialogue history-based methods. (8) DHAP (Ma et al., 2021) uses history memory to store and construct the dynamic query-aware user profile from dialogue history and then uses a personalized decoder to generate a response. Since this model also learns the user profile directly from the dialogue history, it is the most relevant baseline of our method.

4.3 Evaluation

Metric-based. We first evaluate all methods by several metrics with respect to different aspects. (1) BLEU-1/2 (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004) are typical word overlap-based metrics for measuring the similarity between the generated response and the ground-truth.[6] (2) Distinct-1/2 (Li et al., 2016b) consider the number of uni- or bi-grams in the generated response, which is commonly used for evaluating the diversity. (3) The embedding-based metrics (*i.e.*, average, extrema, and greedy) (Chan et al., 2019) are introduced to measure the semantic similarity between the generated response and the ground-truth one. (4) As a personalized dialogue model, following previous studies (Ma et al., 2021), two tailored metrics are adopted to measure how much information is included in the dialogue history that can be reflected in the response. Persona-F1 (P-F1) (Lian et al., 2019) calculates the F1 value to measure the uni-grams co-occurring in both the generated response and the dialogue history. Persona Coverage (P-Cover) (Song et al., 2019) calculates the IDF-weighted word overlap between the generated response and the golden one so that the importance of different words can be taken into account.

Human Annotation. Due to the variability of human language, a response that differs from the ground-truth may also be appropriate. Following previous studies (Chan et al., 2019), we conduct a human evaluation of all methods. Concretely, we sample 100 (query, response, user dialogue history) triplets and hire three well-educated annotators to score the responses generated by different models. Three aspects, *i.e.*, readability, informativeness, and personalization, are considered. The first two factors are scored on a scale of [1, 3] for their quality, while the third is assessed on a scale of [0, 1], indicating whether the response can accurately reflect the user's personality.[7]

4.4 Experimental Results

Metric-based evaluation. Table 1 shows all models' performance under different metrics. On both datasets, it is clear to see that our MSP model outperforms baselines on all metrics. The improvement is statistically significant (t-test with p-value < 0.05). These findings indicate that our model is capable of generating more fluent, diverse, and personalized responses than all baselines. In particular, we can observe: (1) MSP achieves better performance on overlap-based metrics. This suggests that our model can provide responses that are more similar to the ground-truth with the help of the selected tokens. (2) For diversity metrics, the higher distinct values show that our generated responses are more diverse. Additionally, predefined profile-based methods and user dialogue history-based methods outperform others. This shows that incorporating external information can aid in the generation of more informative responses. (3) In addition to generating more overlapped words with the ground-truth response, the improvements of embedding metrics reflect that our model generates more semantically relevant responses. (4) Finally, the increase in personalized metrics implies that our approach can incorporate more user-specific information into the generation. Furthermore, the significant improvement over DHAP demonstrates that our model can extract more meaningful personalized information from the user dialogue history.

Human annotation. The result of human annotation on the Weibo dataset is shown in Table 2. The Fleiss Kappa is around 0.62, indicating a substantial agreement achieved by three annotators. In general, the results of human annotation are consistent with those of the metric-based evaluation. Both of them demonstrate our model's superiority at generating more fluent, informative, and personalized responses. Compared to non-personalized methods, user id-based methods can enhance personalization at the expense of readability. User dialogue history-based methods (*i.e.*, DHAP and MSP) can largely improve the personalization of the response while retaining a high level of readability and informativeness. We attribute this to the abundant personal information contained in the user dialogue history.

[6] The results of BLEU-3/4 and ROUGE-1/2 are provided in Appendix C.

[7] The detailed scoring criteria are described in Appendix D.

Table 1: The result of metric-based evaluation on the Weibo dataset and Reddit Dataset. "†" indicates that our model achieves significant improvement in t-test with p-value < 0.05.

		Overlap-based Metric			Diversity		Embedding Metric			Persona Metric	
		BLEU-1	BLEU-2	ROUGE-L	Dist-1	Dist-2	Average	Extrema	Greedy	P-F1	P-Cover
Weibo	Seq2Seq	3.330†	0.294†	8.985†	0.930†	2.180†	0.321†	0.266†	0.254†	0.154†	0.041†
	MMI	3.631†	0.095†	5.264†	10.710†	43.458†	0.477†	0.696†	0.305†	0.325†	0.054†
	DialoGPT	6.068†	0.741†	8.459†	15.322†	55.536†	0.557†	0.793†	0.324†	0.522†	0.061†
	GPMN	4.899†	0.696†	7.785†	11.724†	32.730†	0.353†	0.391†	0.301†	0.542†	0.084†
	PerCVAE	5.114†	0.299†	7.380†	14.098†	49.733†	0.469†	0.657†	0.299†	0.903†	0.086†
	Speaker	4.994†	0.113†	7.868†	6.035†	19.007†	0.492†	0.712†	0.311†	0.225†	0.082†
	PersonaWAE	3.510†	0.155†	10.546†	2.551†	19.743†	0.563†	0.757†	0.307†	1.740†	0.103†
	DHAP	9.324†	0.894†	14.122†	15.175†	58.806†	0.523†	0.747†	0.313†	1.791†	0.144†
	MSP (Ours)	**11.875**	**5.108**	**15.563**	**24.203**	**73.196**	**0.605**	**0.883**	**0.331**	**2.170**	**0.297**
Reddit	Seq2Seq	1.820†	0.023†	4.069†	5.203†	19.485†	0.545†	0.554†	0.470†	0.051†	0.029†
	MMI	2.065†	0.011†	3.784†	5.914†	31.093†	0.543†	0.607†	0.454†	0.085†	0.038†
	DialoGPT	4.735†	0.397†	8.943†	6.353†	29.106†	0.604†	0.733†	0.448†	0.137†	0.040†
	GPMN	2.686†	0.376†	4.776†	12.325†	35.762†	0.406†	0.331†	0.358†	0.189†	0.037†
	PerCVAE	5.933†	0.576†	8.112†	9.631†	40.213†	0.637†	0.649†	0.499†	0.212†	0.040†
	Speaker	2.642†	0.054†	4.469†	8.951†	34.187†	0.538†	0.606†	0.457†	0.115†	0.031†
	PersonaWAE	2.637†	0.113†	8.199†	1.758†	25.915†	0.629†	0.685†	0.442†	0.206†	0.032†
	DHAP	6.858†	0.737†	11.720†	18.707†	66.932	0.709	0.721†	0.539	0.227†	0.111†
	MSP (Ours)	**7.174**	**0.883**	**12.171**	**21.247**	**68.897**	**0.716**	**0.764**	**0.545**	**0.276**	**0.137**

Table 2: The result of human evaluation on Weibo dataset. "†" indicates that our model achieves significant improvement in t-test with p-value < 0.05.

Model	Readability	Informativeness	Personality
Seq2Seq	1.76†	1.37†	0.11†
MMI	1.96†	1.88†	0.19†
DialoGPT	2.33	2.10†	0.32†
GPMN	2.01†	2.16†	0.35†
PerCVAE	2.10†	2.01†	0.39†
Speaker	1.89†	1.44†	0.24†
PersonaWAE	1.81†	2.01†	0.32†
DHAP	2.29†	2.19†	0.55†
MSP (Ours)	**2.37**	**2.39**	**0.67**
Ground-Truth	2.71	2.66	0.76

Table 3: The results of ablation experiments on Weibo dataset.

Models	BLEU-1	BLEU-2	P-Cover
MSP (Full)	**11.875**	**5.108**	**0.297**
w/o User Refiner	6.093	0.757	0.151
w/o Topic Refiner	6.163	0.839	0.178
w/o Token Refiner	4.213	0.609	0.116
w/o Current U's Profile	9.365	3.146	0.238
w/o Similar Us' Profile	6.413	0.871	0.245
w/o Joint Training	6.070	0.749	0.130

5 Further Analysis

We further conduct a series of analyses to elaborate our model. All analysis here is based on the Weibo dataset, while similar results can be observed on the Reddit dataset.

Ablation Study. To investigate the impact of different modules in MSP, we conduct an ablation study by removing or using different strategies in each module.

We first study the influence of the refiners at three levels. (1) We remove the user refiner and train our model using randomly sampled users. We can see the performance of all metrics drops. This illustrates that our MSP model can select users that shares the same interests as the current user and thereby improving response quality. (2) We remove the topic refiner and supply the token refiner with full dialogue history. The performance degradation demonstrates that various topics in dialogue history introduce lots of noise, misleading the token refiner on extracting valuable tokens, thus impairing the personalized response generation. (3) We eliminate the token refiner and feed all dialogue history sentences directly into the generator.[8] The decline in performance implies the effectiveness and necessity of token selection. It is worth noting that, as compared to using the complete history, our selection strategy can reduce training time by 41.6%, considerably increasing efficiency. All of the aforementioned experimental results suggest that MSP's advantage stems from high-quality personalized information extraction rather than simply introducing additional information.

[8]Due to the length limitation of GPT-2, history with more than 512 tokens will be truncated.

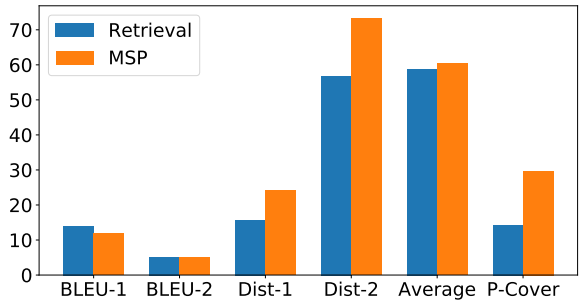

Figure 2: Comparison with the retrieval-based model on the Weibo dataset.[9]

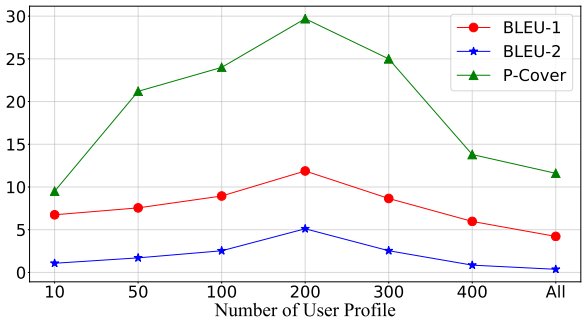

Figure 3: Experiments with the different number of user profiles on the Weibo dataset.[9]

We then explore the impact of personalized information from two sources, *i.e.*, the current user's profile and the similar users' profile. Removing either of them results in decreased performance. This exemplifies their usefulness. Specifically, compared with similar users' profiles, eliminating the current user's profile will hurt the personalization effect heavily. This result shows that, for personalization, the current user's profile is more essential than that of similar users, which is quite intuitive. However, the similar users' profile has a significant effect on BLEU-1/2, implying that such profile can provide abundant information for response generation. Consequently, integrating both types of profiles significantly improves response generation.

Finally, we conduct an experiment to validate our proposed joint training for the token refiner. The declining performance indicates that the token refiner is unable to extract useful information in the absence of additional supervision signals. Indeed, when the sentence matching task is removed, the token refiner extracts tokens that are relevant to the current query, which is less useful for generating a personalized response.

Influence of the selection mechanism. To validate the effectiveness of our proposed selection mechanism, we replace the refiner with a traditional retrieval method (*i.e.*, BM25 (Robertson and Walker, 1994)). Specifically, we use the query to retrieve 15 relevant responses and feed them into our model for training. The experimental results are shown in Figure 2. We can observe that the retrieval strategy achieves comparable performance with our model on word-overlap and embedding-based metrics. This suggests that the relevant dialogue history for the query can provide valuable information for response generation. However, the retrieval strategy performs poorly on diversity and personalization metrics. This demonstrates that, without careful selection, the retrieved information is too generic and thus less helpful for personalized response generation.

Influence of the personalized tokens amount. In MSP, three refiners are designed to extract personalized tokens for response generation. Intuitively, the amount of the tokens will have an effect on the refiner's performance. We report this influence in Figure 3. As we can see in the left part, the quality of response generation improves with more tokens used. This is because fewer tokens are incapable of covering sufficient personalized information for response generation. Our MSP model performs optimally when about 200 personalized tokens are selected. When more tokens are introduced, the performance degrades. The potential reason is that more tokens would bring noise to the generation. This is consistent with our speculation that the dialogue history is noisy and the information selection is both effective and necessary.[10]

6 Conclusion

In this work, we propose an MSP model for personalized response generation. Unlike previous related work, we utilize a refiner structure to extract query-aware persona information from large-scale dialogue history. The multi-level refiners can sparsely extract valuable information from dialogue history and leverage similar users' information to enhance the current user's personalization. Experimental results confirm the effectiveness of our model on generating informative and personalized responses.

[9]To keep the dimensionality consistent, P-Cover is multiplied by a factor of 100.

[10]Due to the space limitation, we present a case study in Appendix E.

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372.

JinYeong Bak and Alice Oh. 2019. Variational hierarchical user-based conversation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1941–1950. Association for Computational Linguistics.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1866–1875. Association for Computational Linguistics.

Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Modeling personalization in continuous space for response generation via augmented wasserstein autoencoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1931–1940. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016c. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087. ijcai.org.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 605–612. ACL.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 555–564. ACM.

Oluwatobi Olabiyi, Anish Khazane, Alan Salimov, and Erik T. Mueller. 2019. An adversarial learning framework for A persona-based multi-turn dialogue model. *CoRR*, abs/1905.01992.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. 2021a. Learning implicit user profiles for personalized retrieval-based chatbot. *CoRR*, abs/2108.07935.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021b. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the SIGIR 2021*. ACM.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4279–4285.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 232–241. ACM/Springer.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5190–5196.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7281–7288. AAAI Press.

Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Chang Yan Chi. 2020. If I hear you correctly: Building and evaluating interview chatbots with active listening skills. *CoRR*, abs/2002.01862.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1341–1350. ACM.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019a. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3721–3730. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664.

Table 4: Statistics of the reddit and weibo datasets.

	Reddit	**Weibo**
# Users	78,031	46,973
Avg. history length	72.385	30.830
Avg. # words of query	19.8	22.9
Avg. # words of response	9.1	9.6
# Training samples	5,734,129	1,495,149
# Validation samples	10,000	10,000
# Testing samples	10,000	10,000

Table 5: The results of extra evaluation metrics on Weibo dataset and Reddit Dataset. "†" indicates that our model achieves significant improvement in t-test with p-value < 0.05. The best results are in bold.

		BLEU-3	BLEU-4	ROUGE-1	ROUGE-2
Weibo	Seq2Seq	0.011†	0.001†	8.740†	0.373†
	MMI	0.046†	0.007†	5.316†	0.105†
	DialoGPT	0.114†	0.027†	9.414†	0.632†
	GPMN	0.359†	0.066†	8.086†	0.753†
	PerCVAE	0.466†	0.089†	7.946†	0.485†
	Speaker	0.107†	0.041†	7.997†	0.155†
	PersonaWAE	0.889†	0.155†	11.341†	0.358†
	DHAP	1.170†	0.401†	14.131†	3.608†
	MSP (Ours)	**3.973**	**3.522**	**16.249**	**5.812**
Reddit	Seq2Seq	0.007†	0.001†	3.989†	0.233†
	MMI	0.007†	0.003†	3.960†	0.245†
	DialoGPT	0.054†	0.010†	8.977†	0.610†
	GPMN	0.039†	0.006†	4.896†	0.330†
	PerCVAE	0.068†	0.009†	8.004†	0.540†
	Speaker	0.021†	0.005†	4.017†	0.245†
	PersonaWAE	0.029†	0.007†	8.247†	0.517†
	DHAP	0.079†	0.013†	10.680†	0.697†
	MSP (Ours)	**0.106**	**0.019**	**11.078**	**0.745**

A Statistics of The Datasets

We adopt our experiment on two different datasets. The Weibo dataset contains about 46K users and about 1.5M samples (query, response, dialogue history), and the Reddit dataset has 78K users and 5.7M samples. The details of statistical information are shown in Table 4.

B Implement Details

We experiment with multiple sets of parameters to select the best model, and the final parameters are as follows: The dimension of the embeddings and Transformer hidden units is 768. The number of heads in the Transformer is 12. We use 12 layers in the decoder and 2 layers in the query encoder. The topic number is 15, and the similar user number is set to 10. The selected profile token number is 200 for the Weibo dataset and 30 for the Reddit dataset. The batch size is 128. Following (Holtzman et al., 2020), we adopt nucleus sampling as our decoding strategy. We use the Adam (Kingma and Ba, 2015) optimizer for refiner and AdamW with a warm-up method for the generator to optimize the parameters in our model and adopt the suggested hyper-parameters for optimization.

Table 6: Criteria of human evaluation.

Metric	Score	Criteria
Read.	1	not a complete sentence or hard to read
	2	grammatically formed
	3	fluent and well to read
Infor.	1	meaningfulness sentence
	2	contains few informative words
	3	have a clear and specific meaning
Per.	0	doesn't resemble any user history
	1	reflect some personal information as same as user history

C Extra Experimental Results

As n-gram word overlap metrics can reflect user speaking style more accurately, we evaluate the BLEU-3/4 (Papineni et al., 2002), and the result is shown in Table 5. It is consistent with other evaluations that our model outperforms every indicator. This demonstrates that user profiles also contain speaking style information, and our model can use the information to achieve a personalized response.

D Criteria of Human Evaluation

Following (Chan et al., 2019), we adopt three aspects to evaluate the generated response. *i.e.*, readability: is the response grammatically formed and smooth; informativeness: does the response contains informative words; personalization: does the response resembles any user history. The details of scoring criteria are shown in Table 6.

E Case Study

To show the effect of our model more concretely, we adopt a case study, and the results are shown in Table 7. It shows that our model can extract profiles from both current and similar users and generate informative and personalized responses. Specifically, in his dialogue history, he mentioned sports **H1, H2** and music **H3, H4** topics. Firstly, we can select similar users who also talk about sports and music, using the user refiner. Then, as the query is related to music topic, we extract the current and sim users' dialogue history responses

Table 7: Case study for one example. Due to space limitations, we present a few user history responses and few references.

History	**H1:** Liverpool's configuration has the life of a champion, support it! **H2:** Champion Liverpool! **H3:** I like to listen to music. **H4:** The songs on this album are really beautiful, so worth enjoying.
Query	New album "How Am I? -The sun rises as it always does", the first single "Ember" heals the system, go listen!
Persona Reference	**R1:** I like to listen to music. **R2:** The songs on this album are really beautiful, so worth enjoying.
Sim Reference	**R1:** Quite like this type of song. **R2:** Angela Leung's "Ember" is really good.
Persona Profile	like, song, album, beautiful, enjoying
Sim Profile	like, song, Angela Leung, good, Ember
Response	**DialoGPT**: My little heart has flown. **MSP(Ours)**: Angela Leung's song is very good. **Golden**: Angela Leung's songs must be listened to.

about music topic **R1, R2** by topic refiner. Furthermore, the token refiner selects some meaningful and personalized words from long sentences of reference. In this case, we can find that the token refiner extracts some compliment words (like, beautiful, enjoying) from current user's history sentence since the current user likes listening to music. And the token refiner captures more concrete tokens from sim users' history sentence, such as "Angela Leung" and "Ember". By combing two profiles, our personalized generator gets an informative and personalized response close to ground-truth. In contrast, DialoGPT generates a fluent but meaningless response to the query.